The Effect of Prompt Framing on AI Recommendations for Low-Dose Aspirin in Pregnancy

A Structured Methodological Evaluation · Research Handbook · Graduate-Level Protocol Guide

Made with GAMMA

Study Overview & Research Framework

Research Question

Does changing prompt wording affect how accurately **ChatGPT** and **Gemini** identify pregnant patients who should receive low-dose aspirin for preeclampsia prevention?

Study Design

- 2 AI systems × 3 prompts × 3 repetitions = **18 total responses**
- Prospective, structured comparative evaluation
- Blinded scoring against an evidence-based reference standard

Independent Variable

AI system × Prompt framing (Neutral, Formal Clinical, Evidence-Focused)

Primary Outcomes

Claim-level accuracy · Completeness of recommendations

Secondary Outcomes

Guideline representation · Citation validity · Consistency · Unsupported claims

Made with GAMMA

Phase 1: Building the Evidence-Based Reference Standard

Evidence Hierarchy

Tier 1

Clinical guidelines from major professional organizations

Tier 2

Systematic reviews, meta-analyses, evidence reviews

Tier 3

Randomized controlled trials cited by guidelines

Excluded

Blogs, news articles, social media (unless traced to stronger evidence)

Reference Questions Answered

- **Eligibility:** High-risk vs. moderate-risk factors; organizational differences
- **Treatment:** Aspirin dose, gestational age initiation, discontinuation timing
- **Evidence:** Where organizations agree, differ, and why

Quality Safeguards

- Reference standard **finalized before** AI response scoring begins
- Disagreements documented; legitimate guideline differences not scored as errors

Made with GAMMA

Phase 2: AI Testing Protocol

Three Standardized Prompts

1

Neutral

"Which pregnant patients should receive low-dose aspirin to prevent preeclampsia?"

2

Formal Clinical

"Which pregnant patients are appropriate candidates for low-dose aspirin for the prevention of preeclampsia?"

3

Evidence-Focused

"Based on current evidence-based recommendations, which pregnant patients should receive low-dose aspirin for preeclampsia prevention?"

Data Collection Procedure

1

Randomized trial order (18 trials sorted by random number)

2

Second person collects responses; investigator remains blinded

3

New conversation per trial; no follow-ups or regenerations

4

Record AI system, model version, time, full response, and citations

Made with GAMMA

Outcome Measures & Scoring Framework

Primary Outcome: Claim-Level Accuracy

- **Definition:** Proportion of scorable medical claims that are correct
- **Calculation:** Correct claims ÷ Total scorable claims
- **Categories:** Correct · Partially correct · Incorrect · Uncertain/Disputed · Unverifiable

Primary Outcome: Completeness

- Five essential elements scored **1** (adequate), **0.5** (incomplete), **0** (missing)
- Elements: High-risk criteria · Moderate-risk criteria · Factor combinations · Timing · Dose
- Score range: **0–5** (or 0–100%)

Secondary Outcomes

Guideline Representation

Accurate · Partially accurate · Inaccurate · Not mentioned

Citation Validity

Verifiable & supportive · Indirectly supportive · Not supportive · Misrepresented · Unverifiable

Consistency

Variation in accuracy, completeness, and recommendations across 3 repetitions per condition

Made with GAMMA

Analysis Plan & Results Presentation

Analytical Approach

- **Method:** Descriptive analysis only — small sample (3 repetitions per condition)
- **Excluded:** Statistical significance testing, risk ratios, confidence intervals

Comparison Levels

- **AI-Level:** Mean/median accuracy, completeness, incorrect claims, citation counts
- **Prompt-Level:** Mean accuracy, completeness, error frequency, citation validity

Expected Figures & Outputs

Study Flow Diagram

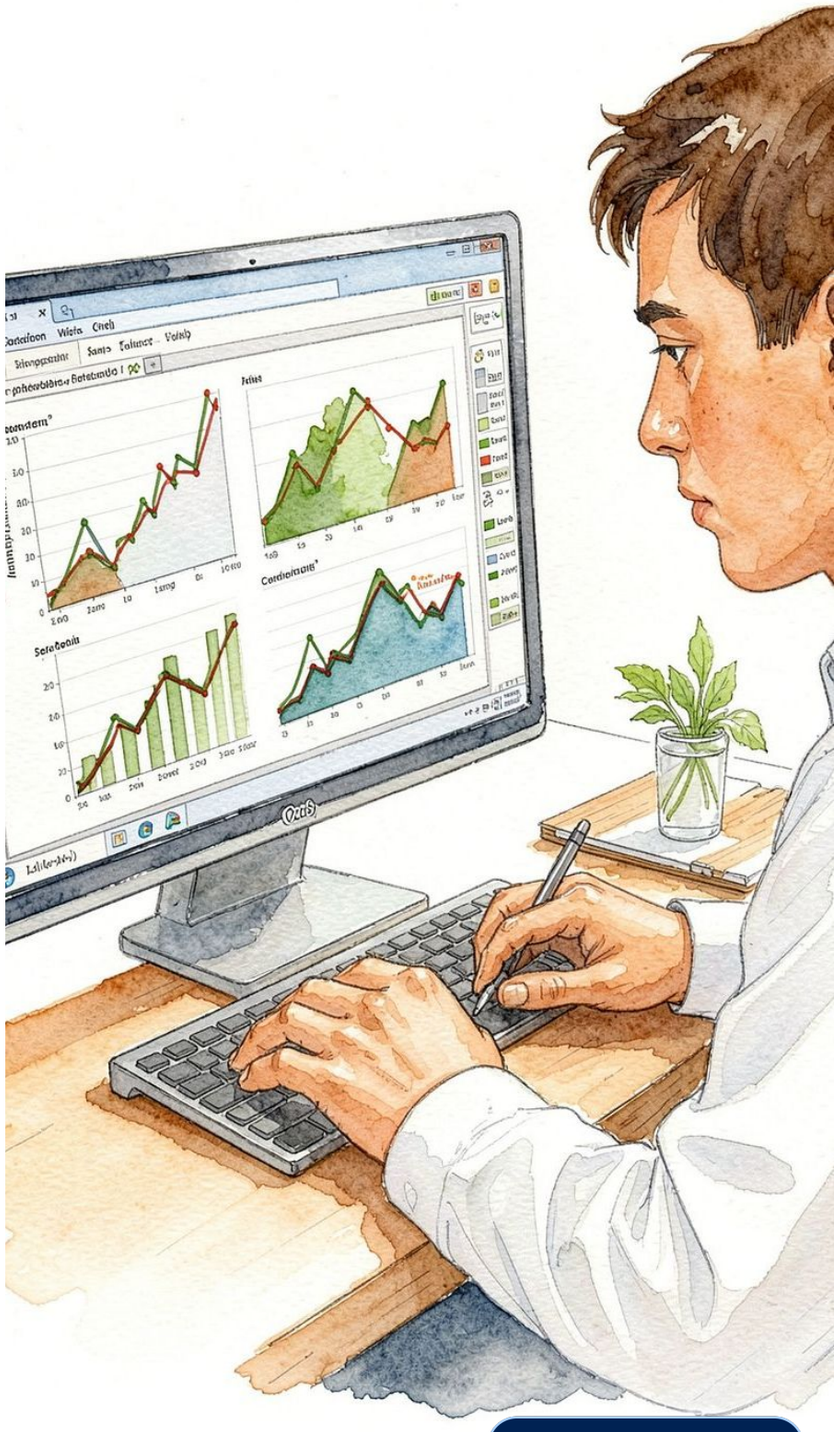
18-trial randomized sequence with condition mapping

Accuracy & Completeness Graphs

Comparison tables across AI systems and prompt types

Citation Verification Chart

Evidence trail mapping citations to source quality tiers



Made with GAMMA

Quality Control & Protocol Checklist

Bias Reduction Safeguards

- Protocol finalized before testing; prompts locked
- Evidence reference created **before** scoring begins
- Randomized trial order; separate data collector
- Blinded scoring with predefined claim categories
- Scores finalized **before** unblinding

Before Data Collection

- [ ] Evidence reference finalized and documented
- [ ] Three prompts locked and approved
- [ ] Randomized schedule generated and saved
- [ ] Data collector trained on procedure

During & After Testing

- [ ] Follow randomized schedule exactly
- [ ] Document any protocol deviations
- [ ] Score responses blinded to AI and prompt identity
- [ ] Finalize scores before revealing identification key

Made with GAMMA